3.2 Stream

2022年3月3日 10:43

MP的问题.

MP属于短任务模式.

完成任务之后进程就Release

不适用于对RT要求高的场景.

↓

实时性任务.

监控类应用.

↑

金融服务

个性化内容实时推荐.

网络监控.

气象测控

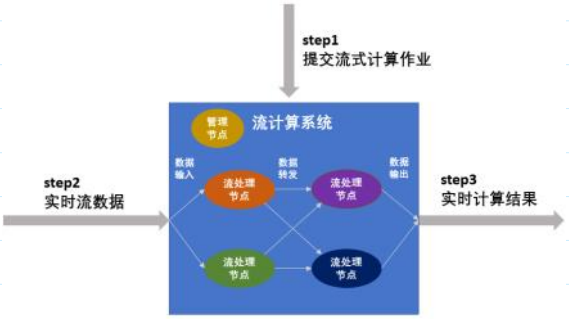
什么是 Stream. → 用于处理数据密集型应用.

流数据：需要进行实时处理的Data.

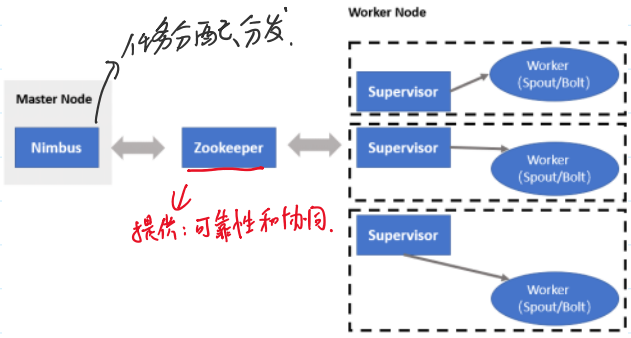
总结来讲，流数据的特征主要包括以下 4 点：

- 数据如流水般持续、快速地到达；
- 海量数据规模，数据量可达到 TB 级甚至 PB 级；
- 对实时性要求高，随着时间流逝，数据的价值会大幅降低；
- 数据顺序无法保证，也就是说系统无法控制将要处理的数据元素的顺序。

Stream工作原理.

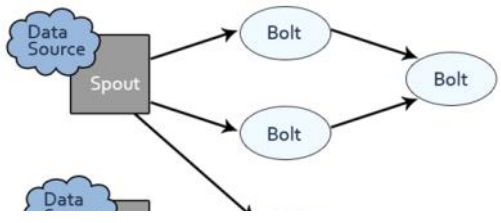


Storm的工作原理.



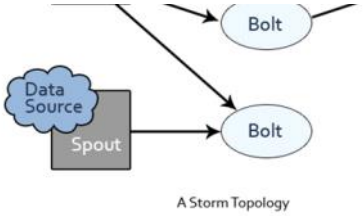
在详细介绍 Worker 组件之前，我首先介绍一下 Storm 的核心抽象：数据流。数据流是一个无界序列，是在分布式环境中并行创建、处理的一组元组（tuple）。数据流可以由一种能够表述数据流中元组的域（fields）的模式来定义。

Storm数据流转换组件.



- Spout 用于接收源数据。通常情况下，Spout 会从一个外部的数据源读取数据元组，然后将它们发送到拓扑中。例如，Spout 从 Twitter API 读取推文并将其发布到拓扑中。

- Bolt 负责处理输入的数据流，比如数据过滤（filtering）、函数处理（functions）、聚



A Storm Topology

- Bolt 负责处理输入的数据流，比如数据过滤（filtering）、函数处理（functions）、聚合（aggregations）、联结（joins）、数据库交互等。数据处理后可能输出新的流作为下一个 Bolt 的输入。每个 Bolt 往往只具备单一的计算逻辑。当我们执行简单的数据流转换时，比如仅进行数据过滤，则通常一个 Bolt 可以实现；而复杂的数据流转换通常需要使用多个 Bolt 并通过多个步骤完成，比如在神经网络中，对原始数据进行特征转换，需要经过数据过滤、清洗、聚类、正则化等操作。

批量计算 VS. 流计算

对比指标	批量计算	流式计算
数据特点	大规模非实时数据	持续产生的、具有易逝性的实时数据
数据集成方式	预先加载并存储批量的数据	实时加载数据，不存储数据
数据处理方式	对集中的所有数据或大批量数据一起进行处理	对最近输入的数据进行处理
计算规则	1. 任务处理过程中计算逻辑可以修改 2. 计算逻辑修改后，数据可重新计算	1. 任务处理过程中计算逻辑不可以修改 2. 计算逻辑一旦修改，之前的数据不可重新计算
实时性分析	处理需要几分钟甚至是几小时	处理时间仅为几秒或者几毫秒
适用场景	用于对时延不敏感的批处理任务，比如大规模信息统计等	用于对时延敏感的数据密集型任务，该任务可拆分为小批量数据的任务，如实时天气预报、直播中音视频流处理等
典型计算框架	MapReduce等	Storm、InfoSphere Streams、Dstream等